Technical Appendix for Fast Redescription Mining Using Locality-Sensitive Hashing

Anonymous Author(s)

No Institute Given

1 Experimental Evaluation

1.1 Dataset Properties

Properties of all data sets used in the experiments are presented in Table 1.

1.2 Finding Initial Pairs

We compared the accuracy of the initial pairs found by `ReReMiBkt` and `ReReMi`. The results are in Fig. 1 for the CMS_{116395}, CMS_{7199}, and MammalsW datasets. They show that `ReReMiBkt` and `ReReMi` get the same results with the MammalsW dataset and there are no dots in the x-axis. With the CMS data sets, `ReReMi` and `ReReMiBkt` find somewhat different initial pairs. `ReReMiBkt` finds almost always equally good initial pairs as `ReReMi`, as well as many intial pairs that `ReReMi` does not find. This indicates that comparison to `ReReMiBkt` for the quality of the initial pairs is fair.

Sensitivity to parameters. We evaluated the sensitivity of `Fier_init` to the parameters affecting locality-sensitive hashing. Overall, we found it to be very robust.

We tested all combinations of values 20, 40, 60, and 80 for b_J and values 3, 5, 7, 10, 12, and 15 for r_J and evaluated the results both w.r.t. running time and quality of answers. For these experiments we used the EuroClim data set.

Table 1. Data set properties.

data set	$\lvert\mathcal{E}\rvert$, entities	$\mathbf{D}_L$ attributes	$\mathbf{D}_R$ attributes
EuroClim	2 575	12 numerical	12 numerical
NoisyClim	2 575	36 numerical	36 numerical
MammalsW	54 013	48 numerical	4754 Boolean
Ethno	1 267	23 numerical and categorical	90 numerical
DentalW	28 886	11 numerical	19 numerical
DentalA	6 404	7 numerical	19 numerical
VAA	1 656	9 categorical	107 categorical
CMS_d	d	152 numerical	452 numerical

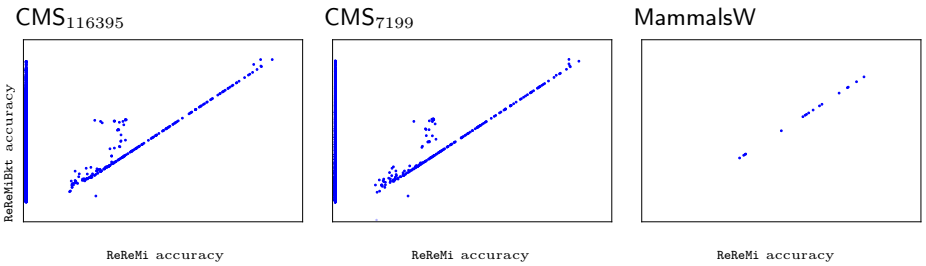

Fig. 1. Comparing the accuracy of pairs found by `ReReMiBkt` and `ReReMi`. Each dot represents a pair of columns, and its location indicates the highest-accuracy initial pair `ReReMiBkt` and `ReReMi`.

Table 2. The effect of LSH parameters to running time (left) and average Jaccard of the fifth best result $\bar{J}$@5 (right) depending on the number (b) and the width (r) of bands in LSH, for five repetitions and using EuroClim data.

	time (s)						$\bar{J}$@5					
	r_J						r_J					
b_J	3	5	7	10	12	15	3	5	7	10	12	15
20	33.89	21.34	15.00	6.94	5.68	5.14	0.69	0.70	0.63	0.69	0.50	0.63
40	45.71	33.22	23.09	13.46	11.30	9.54	0.71	0.73	0.65	0.72	0.57	0.67
60	56.91	40.77	29.79	18.59	15.66	13.88	0.73	0.73	0.65	0.71	0.62	0.70
80	68.86	48.69	35.61	23.76	20.35	17.73	0.72	0.71	0.66	0.73	0.63	0.70

Table 2 shows how the parameters impact the running time and accuracy. We can see that r_J has the largest impact, the running time increasing with smaller values of r_J. This is because r_J determines the length of the minhash signature, and shorter signatures imply a higher chance of matching and forming a candidate pair. On the other hand, the parameters do not have much impact on the average accuracy of the fifth-best result ($\bar{J}$@5).

As the average quality does not change by much, we can conclude that the user can set rows and bands primarily based on how many initial pairs they want to find, but that the settings we have used throughout these experiments ($r_J = 10$, $b_J = 40$) seem to give a good balance.

1.3 Extending Initial Pairs

Extending multiple times. The full results for extending pre-mined initial pairs are presented in Table 3.

1.4 Building Full Redescriptions

The results are presented in Table 4.

Table 3. Accuracy, total number of extensions, and time in seconds when extending the `ReReMi` initial pairs multiple times. All results are averages over 5 runs.

			MammalsW			DentalW		
alg.	r_H	b_H	J@10	# of ext.	time (s)	J@10	# of ext.	time (s)
`Fier_ext`	10	10	0.77	296.4	1471.74	0.60	535.8	643.16
		20	0.79	329.0	1664.60	0.61	547.8	712.50
		40	0.79	354.8	1764.90	0.62	543.8	776.40
		80	0.82	447.8	2176.34	0.61	545.6	819.52
	20	10	0.72	112.6	768.05	0.57	500.2	390.47
		20	0.72	130.6	973.91	0.58	523.0	472.29
		40	0.75	237.4	1383.91	0.59	530.6	578.33
		80	0.75	259.6	1674.15	0.60	541.0	727.70
	30	10	0.68	60.0	438.41	0.54	402.8	245.93
		20	0.70	65.4	641.48	0.57	462.2	348.41
		40	0.71	101.0	990.65	0.57	512.8	483.04
		80	0.72	133.8	1369.42	0.58	519.0	686.52
	40	10	0.63	38.4	285.89	0.51	229.0	125.20
		20	0.68	54.6	473.24	0.53	329.0	221.54
		40	0.70	62.0	727.29	0.54	426.2	395.37
		80	0.70	63.6	1162.73	0.57	480.0	658.67
	50	10	0.62	34.6	209.51	0.48	105.4	79.45
		20	0.63	40.8	333.93	0.50	168.4	133.86
		40	0.67	53.8	677.99	0.52	255.0	256.28
		80	0.69	57.2	1122.53	0.54	349.8	538.83
`ReReMi`			0.83	410.0	3263.28	0.61	517.0	1494.44

			CMS_{7199}			CMS_{116395}		
alg.	r_H	b_H	J@10	# of ext.	time (s)	J@10	# of ext.	time (s)
`Fier_ext`	10	10	0.82	28.8	90.08	0.77	90.0	1072.10
		20	0.82	40.0	181.25	0.78	106.8	1906.59
		40	0.82	47.0	299.37	0.80	126.4	2646.13
		80	0.82	52.4	383.05	0.80	127.6	4366.60
	20	10	0.80	4.0	7.56	0.72	11.4	202.65
		20	0.81	6.4	17.68	0.73	20.4	303.56
		40	0.81	8.6	31.13	0.74	41.8	532.56
		80	0.82	14.6	60.59	0.75	55.0	964.70
	30	10	0.80	0.6	6.82	0.71	5.0	148.69
		20	0.80	1.4	12.58	0.72	5.6	228.74
		40	0.81	3.0	23.88	0.72	8.4	421.66
		80	0.81	4.6	46.21	0.72	9.4	730.61
	40	10	0.79	0.0	7.35	0.71	1.4	117.36
		20	0.79	0.6	13.86	0.71	2.8	223.92
		40	0.80	0.8	25.98	0.71	2.8	396.22
		80	0.80	0.8	49.29	0.71	4.6	732.58
	50	10	0.79	0.0	7.98	0.71	0.6	120.49
		20	0.79	0.0	14.90	0.71	0.8	197.45
		40	0.79	0.4	28.25	0.71	1.6	368.98
		80	0.80	0.8	55.20	0.71	2.8	730.67
`ReReMi`			0.83	55.0	695.85	0.81	154.0	7626.95

Table 4. Accuracy, total number of extensions, total number of resulting redescriptions, and total time in seconds for full redescription mining. All results are averages over 5 runs.

			MammalsW				DentalW			
alg.	r_H	b_H	J@10	# ext.	# results	time	J@10	# ext.	# results	time
`Fier_full`	20	10	0.68	60.4	19.2	518	0.55	446.8	194.6	345
		20	0.74	94.6	32.4	765	0.57	486.8	216.8	445
		40	0.72	75.0	24.0	901	0.59	505.8	222.6	549
	30	10	0.60	33.2	12.4	285	0.53	348.2	169.0	217
		20	0.63	39.0	14.2	446	0.52	423.8	189.0	315
		40	0.69	66.8	20.8	705	0.55	454.6	202.6	447
	40	10	0.56	28.2	13.0	213	0.48	227.6	124.0	144
		20	0.63	28.0	10.6	333	0.51	258.8	139.6	199
		40	0.58	37.0	13.2	500	0.53	362.8	178.2	358
`ReReMi`			0.83	410.0	71.0	3402	0.61	517.0	93.0	3063
`ReReMiBkt`			0.83	410.0	71.0	3381	0.63	512.0	90.0	1866

			CMS_{7199}				CMS_{116395}			
alg.	r_H	b_H	J@10	# ext.	# results	time	J@10	# of ext.	# results	time
`Fier_full`	20	10	0.74	8.8	18.4	30	0.74	9.6	19.2	352
		20	0.75	12.2	18.8	34	0.76	19.6	21.4	463
		40	0.77	23.2	23.0	53	0.76	22.8	23.8	592
	30	10	0.74	2.0	19.8	24	0.73	1.6	17.6	302
		20	0.74	3.6	17.0	30	0.75	4.2	17.2	381
		40	0.75	8.8	19.8	41	0.75	6.4	18.4	517
	40	10	0.73	0.6	15.0	24	0.72	1.0	15.4	295
		20	0.73	3.0	17.6	31	0.73	0.6	16.0	367
		40	0.72	2.4	17.2	41	0.74	1.4	16.4	510
`ReReMi`			0.83	55.0	22.0	1032	0.81	154.0	45.0	9588
`ReReMiBkt`			0.83	55.0	22.0	957	0.83	79.0	29.0	5797

			Ethno				VAA			
alg.	r_H	b_H	J@10	# ext.	# results	time	J@10	# of ext.	# results	time
`Fier_full`	20	10	0.55	79.2	50.2	38	0.00	0.0	0.0	0
		20	0.55	105.6	55.8	53	0.39	0.0	1.0	2
		40	0.55	137.4	66.2	74	0.00	0.0	0.0	0
	30	10	0.49	30.0	37.0	15	0.00	0.0	0.0	0
		20	0.51	39.4	34.8	22	0.00	0.0	0.0	0
		40	0.55	69.6	45.4	34	0.00	0.0	0.0	0
	40	10	0.48	6.0	27.8	10	0.00	0.0	0.0	0
		20	0.45	13.0	25.0	12	0.36	0.0	1.0	2
		40	0.50	26.6	35.2	18	0.00	0.0	0.0	0
`ReReMi`			0.65	13.0	3.0	184	0.39	6.0	4.0	8